Optimality Functions in Stochastic Programming

J.O. Royset[*]

Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA

December 2, 2009

Abstract. Optimality functions in nonlinear programming conveniently measure, in some sense, the distance between a candidate solution and a stationary point. They may also provide guidance towards the development of implementable algorithms. In this paper, we use an optimality function to construct procedures for validation analysis in stochastic programs with nonlinear, possibly nonconvex, expected value functions as both objective and constraint functions. We construct an estimator of the optimality function and examine its consistency, bias, and asymptotic distribution. The estimator leads to confidence intervals for the value of the optimality function at a candidate solution and, hence, provides a quantitative measure of solution quality. We also construct an implementable algorithm for solving smooth stochastic programs based on sample average approximations and the optimality function estimator. Preliminary numerical tests illustrate the proposed algorithm and validation analysis procedures.

Keywords: Stochastic programming; nonlinear programming; optimality conditions; validation analysis.

1 Introduction

Stochastic optimization problems arise in numerous context where decisions must be made in the presence of data uncertainty; see the books [13, 9, 20, 17, 37, 34] and references therein for algorithms, models, and applications. In this paper, we deal with a class of stochastic optimization problems defined in terms of expected values of random functions. Let $F^j : \mathbb{R}^n \times \Omega \to \mathbb{R}$, $j = 0, 1, 2, ..., q$, be random functions defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with $\Omega \subset \mathbb{R}^d$ and $\mathcal{F} \subset 2^\Omega$ being the Borel sigma algebra. Moreover, let the expected value functions $f^j : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ be defined by

$$f^j(x) \overset{\triangle}{=} E[F^j(x, w)] \tag{1}$$

for all $j \in \mathbf{q}_0 \overset{\triangle}{=} \{0\} \cup \mathbf{q}$, with $\mathbf{q} \overset{\triangle}{=} \{1, 2, ..., q\}$, where E is the expectation with respect to $\mathcal{P}$. Below we impose conditions that ensure finiteness of $f^j(x)$, $j \in \mathbf{q}_0$, for all $x \in \mathbb{R}^n$ of interest. Optimization problems involving such expected value functions are generally challenging to solve due to the need for estimating expectations repeatedly during the optimization. Even assessing how "close" a given

[*] Tel.: + 1 831 656 2578, fax: +1 831 656 2595, email joroyset@nps.edu.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **02 DEC 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **Optimality Functions in Stochastic Programming**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Operations Research Department,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
in review

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**30**	

candidate point $x \in \mathbb{R}^n$ is to optimality or stationarity may be nontrivial; see [4] and the review below. We specifically consider the stochastic optimization problem

$$P: \quad \min_{x \in \mathbb{R}^n} \{f^0(x) \mid f^j(x) \leq 0, j \in \mathbf{q}\}, \tag{2}$$

where we assume that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are continuously differentiable, but possibly nonconvex. Nonconvex stochastic optimization problems arise in such diverse applications as estimation of mixed logit models [2], engineering design [29], and inventory control [39]. We focus on the assessment of a candidate point $x \in \mathbb{R}^n$ for P, which we refer to as validation analysis, but also consider the generation of such candidate points by an algorithm.

Stationary points of P are defined by the Karush-Kuhn-Tucker (KKT) or the Fritz-John first-order necessary optimality conditions; see for example Propositions 3.3.1 and 3.3.5 in [7] or Theorem 2.2.4 in [25]. If the evaluation of $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$, could be accomplished in finite (and relatively short) time, then it would be possible to verify whether $x \in \mathbb{R}^n$ is stationary or near-stationary by solving a convex quadratic program, see, e.g, Theorem 2.2.8 in [25]. In the present context, however, $f^j(\cdot)$, $j \in \mathbf{q}_0$, are defined in terms of expectations and, as we will see, $\nabla f^j(x)$, $j \in \mathbf{q}_0$, are defined similarly. Hence, $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$, cannot generally be evaluated in finite time resulting in challenging validation analysis for P.

Previous studies of validation analysis in stochastic programming often deal with special cases of P. In the case of uncertainty in the objective function only, i.e., $F^j(\cdot, \omega)$, $j \in \mathbf{q}$, are deterministic functions, [24] and [22] propose procedures for estimating upper and lower bounds on the optimal value of P. Other efforts to compute bounds on the optimal value include [15, 3]; see also the tutorial [4]. These procedures are limited to convex problems as they require global minima of sample average approximations constructed by replacing the expectations in P by their sample averages, or as they make use of strong duality.

Under the same assumption of deterministic constraints, [35] develops confidence regions for $\nabla f(x)$ as well as hypothesis tests for whether a point $x \in \mathbb{R}^n$ satisfies the KKT conditions; see also [14]. The results in [35] can be extended to constraints defined in terms of expectations [33], though that result appears unpublished. The hypothesis tests require that the gradients of the active constraints are linearly independent, the strict complimentary condition holds at x, and that the inverse of an estimate of a variance-covariance matrix is nonsingular.

Stochastic programs with chance constraints are outside the scope of this paper (see for example [21] and Chapter 4 of [34]), but relate to P as they are essentially of the same form as P except that $F^0(\cdot, \omega)$ is deterministic and $F^j(\cdot, \omega)$, $j \in \mathbf{q}$, are indicator functions. For such programs validation analysis may involve estimating the probability of feasibility for a candidate point as well as the use of Lagrangian relaxation to obtain bounds on the optimal value; see for example Section 5.7 in [34] and references therein.

Validation analysis for the full problem P has received less attention. On p. 208 in [34], a lower bound on the optimal value of P is proposed based on the Lagrangian function. The bound, however, may be rather weak in the case of a nonconvex problem. Section 5.2 of [34] (see also [32, 12, 2]) uses stochastic variational inequalities to analyze optimality conditions for P. The results include conditions for almost sure convergence of stationary points of sample average approximations to stationary points of P as the sample size grows. Extension of such results to second-order optimality conditions are found in [2]. A similar result for the case with a nonsmooth objective function and deterministic constraints is found in [39].

In Section 5.2 of [34], we find that under the linear independence constraint qualification and the strict complementarity condition, a stationary point of a sample average approximation with sample size N is approximately normally distributed with mean equal to a stationary point of P and with standard deviation proportional to $N^{-1/2}$. In [30] (see also [31]), we find a hypothesis test for checking whether a candidate solution and a corresponding Lagrange multiplier vector satisfy the KKT conditions for the case with both inequality and equalities defined in terms of expectations. That paper also presents confidence intervals for the constraint functions at a candidate point. While these results are important, they do not directly quantify the quality of a candidate solution that is not stationary for a sample average approximation. In practice, we can usually only hope for a near-stationary solution of P and its sample average approximations. Hence, it becomes important to assess the quality of such solutions.

In this paper, we develop procedures for validation analysis of a candidate point $x \in \mathbb{R}^n$. Since P may be nonconvex, we focus on first-order necessary optimality conditions as validation analysis by means of bounds on the optimal value (see [4]) appears difficult. Specifically, as in Section 2.2.1 of [25], we consider Fritz-John conditions for P as characterized by a nonpositive *optimality function* which vanishes at feasible stationary points. Hence, the value of the optimality function at x gives a measure of the quality of x. We specifically provide bounds on the distance between $f^0(x)$ and the optimal value of P in terms of the value of the optimality function at x.

The optimality function involves $f^j(\cdot)$ and $\nabla f^j(\cdot)$, $j \in \mathbf{q}_0$, and can therefore only be estimated. We develop a strongly consistent estimator for the optimality function at x and examine its asymptotic distribution and bias. We also develop procedures for estimating probabilistic lower bounds on the optimality function at x and corresponding confidence intervals. Since the optimality function is nonpositive and vanishes at a feasible stationary point, such a lower bound provides a conservative estimate of the quality of x. The lower bounds may also lead to a stopping criterion for algorithms for solving P. In contrast to the hypothesis tests in [35], which check the KKT conditions and require linear independence and strict complementary constraint qualifications, we adopt the more general Fritz-John conditions and require no constraint qualification.

The estimator of the optimality function can be viewed as an optimality function of a sample average approximation of P in the case $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, are continuously differentiable for almost all $\omega \in \Omega$. We exploit this observation and develop a convergent algorithm for P under this additional assumption and a constraint qualification.

In Section 2, we define optimality conditions for P in terms of an optimality function and show how that function measures the distance to the optimal value of P. Section 3 constructs an estimator for the optimality function, proves its consistency, and derives the asymptotic distribution of an appropriately scaled and shifted estimator. Section 4 develops procedures for validation analysis by means of the estimator of the optimality function. Section 5 constructs consistent approximations and presents an algorithm for P. Section 6 gives illustrative numerical examples.

2 Optimality Function

In this section we state optimality conditions for P, define an optimality function, and prove a relationship between the optimality function at a feasible point $x \in \mathbb{R}^n$ and the distance between $f^0(x)$ and the optimal value of P. We adopt the Fritz-John first-order necessary optimality conditions; see for example Theorem 2.2.4 in [25]. Before we state the conditions, we give assumptions which ensure that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are finite valued and continuously differentiable. We observe that since $F^j(\cdot,\cdot)$, $j \in \mathbf{q}_0$, are random functions, it follows by definition that $F^j(x,\cdot)$, $j \in \mathbf{q}_0$, are measurable for every $x \in \mathbb{R}^n$.

Assumption 1 *We assume that for a given set $S \subset \mathbb{R}^n$, the following hold for any nonempty compact set $X \subset S$ and for all $j \in \mathbf{q}_0$:*

(i) *There exists a measurable function $C : \Omega \to [0,\infty)$ such that $E[C(\omega)] < \infty$ and for all $x \in X$ and almost every $\omega \in \Omega$, $|F^j(x,\omega)| \le C(\omega)$.*

(ii) *There exists a measurable function $L : \Omega \to [0,\infty)$ such that $E[L(\omega)] < \infty$ and*

$$|F^j(x,\omega) - F^j(x',\omega)| \le L(\omega)\|x - x'\| \tag{3}$$

for all $x, x' \in S$ and almost every $\omega \in \Omega$.

(iii) *For every $x \in X$, $F^j(\cdot,\omega)$ is continuously differentiable at x for almost all $\omega \in \Omega$.*

Assumption 1 is rather weak and commonly made in the literature; see for example Theorem 7.52 in [34]. A large number of applications satisfy Assumption 1 including many instances of two-stage stochastic programs with recourse [17], Conditional Value-at-Risk problems [27], inventory control problems [39], and engineering design problems [29]. If Assumption 1 holds on an open set S and $X \subset S$ is compact, then it follows from Theorem 7.52 in [34] that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are continuously differentiable on X and that $\nabla f^j(x) = E[\nabla_x F^j(x,\omega)]$, for all $x \in X$ and $j \in \mathbf{q}_0$.

We need the following notation. For any vector v, we adopt the convention that $v^j \in \mathbb{R}$ denotes the vector's j-th component. Let

$$\Sigma_q^0 \triangleq \left\{ \mu \in \mathbb{R}^{q+1} \;\middle|\; \sum_{j \in \mathbf{q}_0} \mu^j = 1, \mu^j \ge 0, j \in \mathbf{q}_0 \right\}, \tag{4}$$

$\psi(x) \overset{\triangle}{=} \max_{j\in\mathbf{q}} f^j(x)$, and $\psi(x)_+ \overset{\triangle}{=} \max\{0, \psi(x)\}$.

The Fritz-John first-order necessary conditions for P takes the following form; see, for example, Theorem 2.2.4 in [25].

Proposition 1 *If $\hat{x} \in \mathbb{R}^n$ is a local minimizer for P and Assumption 1 holds on an open set $S \subset \mathbb{R}^n$ containing $\hat{x}$, then there exists a multiplier vector $\hat{\mu} \in \Sigma_q^0$ such that*

$$\sum_{j\in\mathbf{q}_0} \hat{\mu}^j \nabla f^j(\hat{x}) = 0 \tag{5}$$

and

$$\sum_{j\in\mathbf{q}} \hat{\mu}^j f^j(\hat{x}) = 0. \tag{6}$$

□

We refer to a point $\hat{x} \in \mathbb{R}^n$ that satisfies (5) and (6) for some $\hat{\mu} \in \Sigma_q^0$ as a Fritz-John point. We remark that the Fritz-John conditions reduce to the KKT conditions when $\hat{\mu}^0 > 0$; see [25], p. 189.

We follow [25], see p. 190, and express the Fritz-John conditions by means of a continuous optimality function $\theta : \mathbb{R}^n \to (-\infty, 0]$ defined by

$$\begin{aligned}\theta(x) \quad &\overset{\triangle}{=} \quad \min_{h\in\mathbb{R}^n} \Bigg\{ \max\Big\{ -\psi(x)_+ + \langle \nabla f^0(x), h\rangle, \\ &\qquad \max_{j\in\mathbf{q}}\{f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle\}\Big\} + \tfrac{1}{2}\|h\|^2 \Bigg\}.\end{aligned} \tag{7}$$

We find the following alternative expression for $\theta(x)$ useful; see Theorem 2.2.8 in [25]:

$$\theta(x) = -\min_{\mu\in\Sigma_q^0} \Big\{ \mu^0\psi(x)_+ + \sum_{j\in\mathbf{q}} \mu^j[\psi(x)_+ - f^j(x)] + \tfrac{1}{2}\Big\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j(x) \Big\|^2 \Big\}. \tag{8}$$

Let $X_\psi \overset{\triangle}{=} \{x \in \mathbb{R}^n \mid \psi(x) \le 0\}$ be the feasible region of P. The optimality function equivalently expresses the Fritz-John conditions in the sense stated next; see Theorem 2.2.8 in [25].

Proposition 2 *Suppose that $\hat{x} \in X_\psi$ and Assumption 1 holds on an open set $S \subset \mathbb{R}^n$ containing $\hat{x}$. Then, $\theta(\hat{x}) = 0$ if and only if there exists a multiplier vector $\hat{\mu} \in \Sigma_q^0$ such that (5) and (6) hold.*

In view of Proposition 2, the closeness of $\theta(x)$ to zero indicates the proximity of x to a Fritz-John point. Under a convexity assumption, $\theta(x)$ also gives a bound on the distance between $f^0(x)$ and the minimum value of P as the next result shows. We find a similar result for two-stage stochastic program with recourse in [14].

Proposition 3 *Suppose that X_ψ is nonempty, $f^j(\cdot)$, $j \in \mathbf{q}_0$, are twice continuously differentiable, and that there exist constants $0 < m \le 1 \le M < \infty$ such that*

$$m\|x' - x\| \le \langle x' - x, \nabla^2 f^j(x)(x' - x)\rangle \le M\|x' - x\|, \tag{9}$$

for all $x, x' \in \mathbb{R}^n$ and $j \in \mathbf{q}_0$. Then, there exists a constant $c < \infty$ such that for any $x \in X_\psi$,

$$\frac{\theta(x) - c\sqrt{-\theta(x)}}{m} \leq f^0(\hat{x}) - f^0(x) \leq \theta(x)/M, \tag{10}$$

where $\hat{x} \in \mathbb{R}^n$ is the optimal solution of P.

Proof: See Appendix. □

In view of the above results, the optimality function offers a convenient way of measuring the quality of a candidate point. Moreover, as we see in Section 5, the optimality function also provides guidance towards the construction of an implementable algorithm for P.

The computation of $\theta(x)$ for a given $x \in \mathbb{R}^n$ requires the solution of a convex quadratic program with linear constraints (see (8)), which can be achieved in finite time. However, the definition of $\theta(x)$ involves $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$, that, in general, cannot be computed in finite time. Consequently, we define an estimator for $\theta(x)$ using the sample average estimators for $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$.

3 Estimator of Optimality Function

3.1 Definition and Consistency

Let $\omega_1, \omega_2, ...$ be an infinite sequence of independent random vectors each with value in Ω and distributed as $\mathcal{P}$. Let $\mathbb{N} \stackrel{\triangle}{=} \{1, 2, 3, ...\}$. We define for any $N \in \mathbb{N}$, $j \in \mathbf{q}_0$, and $x \in \mathbb{R}^n$, the estimators for $f^j(x)$, $\nabla f^j(x)$, $\psi(x)$, and $\psi(x)_+$ by

$$f^j_N(x) \stackrel{\triangle}{=} \frac{1}{N}\sum_{i=1}^{N} F(x, \omega_i), \tag{11}$$

$$\nabla f^j_N(x) \stackrel{\triangle}{=} \frac{1}{N}\sum_{i=1}^{N} \nabla_x F(x, \omega_i), \tag{12}$$

$\psi_N(x) \stackrel{\triangle}{=} \max_{j \in \mathbf{q}} f^j_N(x)$, and $\psi_N(x)_+ \stackrel{\triangle}{=} \max\{0, \psi_N(x)\}$, respectively. We refer to [11] for an overview of alternative approaches to estimating $\nabla f^j(x)$. In some situations it may be possible to use variance reduction techniques to define alternative estimators with smaller variance than those defined above; see for example Section 5.5 in [34]. However, such estimators are beyond the scope of the paper.

Finally, we define the estimator of $\theta(x)$ by

$$\begin{aligned}\theta_N(x) \quad \stackrel{\triangle}{=} \quad \min_{h \in \mathbb{R}^n} \Bigg\{ & \max\Big\{ -\psi_N(x)_+ + \langle \nabla f^0_N(x), h\rangle, \\ & \max_{j \in \mathbf{q}}\{f^j_N(x) - \psi_N(x)_+ + \langle \nabla f^j_N(x), h\rangle\}\Big\} + \tfrac{1}{2}\|h\|^2 \Bigg\}.\end{aligned} \tag{13}$$

As commonly done, we view $f_N^j(x)$, $j \in \mathbf{q}_0$, $\psi_N(x)$, $\psi_N(x)_+$, and $\theta_N(x)$ as random variables and $\nabla f_N^j(x)$, $j \in \mathbf{q}_0$, as random vectors defined on the product space generated by $(\Omega, \mathcal{F}, \mathcal{P})$ and denote the resulting probability measure and sample space by $\overline{\mathcal{P}}$ and $\overline{\Omega}$, respectively; see Chapter 7 of [34] for further background. To emphasize the dependence on $\overline{\omega} \overset{\triangle}{=} (\omega_1, \omega_2, ...)$ we occasionally write $f_N^j(x, \overline{\omega})$, $\psi_N(x, \overline{\omega})$, etc. Usually, however, we omit $\overline{\omega}$. We observe that under Assumption 1, for all $x \in X$, $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, are continuously differentiable at x for almost every $\omega \in \Omega$. Hence, $\nabla f_N^j(x, \overline{\omega})$, $j \in \mathbf{q}_0$, and $\theta_N(x, \overline{\omega})$ are defined for almost every $\overline{\omega} \in \overline{\Omega}$.

Similar to (8), we deduce from Theorem 2.2.8 of [25] the following equivalent and useful expression for $\theta_N(x)$:

$$\theta_N(x) = -\min_{\mu \in \Sigma_q^0} \Big\{ \mu^0 \psi_N(x)_+ + \sum_{j \in \mathbf{q}} \mu^j [\psi_N(x)_+ - f_N^j(x)] + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f_N^j(x) \Big\|^2 \Big\}. \tag{14}$$

We observe that since the objective function in (8) is a function of expectations, standard results about the relationship between the optimal value of a problem and those of its sample average approximations (see for example Chapter 5 in [34]) are not applicable. In the following, however, we make use of similar proof techniques as in Chapter 5 of [34].

The next result proves that $\theta_N(x)$ is a strongly consistent estimator of $\theta(x)$. This result is similar to classic results about almost sure convergence of optimal values of sample average approximations to the optimal value of an original problem; see, e.g., [19, 26].

Theorem 1 *Suppose that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$. Then, $\theta_N(x) \to \theta(x)$, as $N \to \infty$, almost surely.*

Proof: It follows from Assumption 1 that for all $j \in \mathbf{q}$, $f^j(x)$ is well defined and finite valued and hence the strong law of large numbers implies that $f_N^j(x) \to f^j(x)$, as $N \to \infty$, almost surely. Moreover, Theorem 7.52 in [34] gives that $\nabla f_N^j(x) \to \nabla f^j(x)$, as $N \to \infty$, almost surely. We define for any $\mu \in \Sigma_q^0$ the function $\eta : \Sigma_q^0 \to \mathbb{R}$ by

$$\eta(\mu) \overset{\triangle}{=} \mu^0 \psi(x)_+ + \sum_{j \in \mathbf{q}} \mu^j [\psi(x)_+ - f^j(x)] + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j(x) \Big\|^2 \tag{15}$$

and similarly we define the function $\eta_N : \Sigma_q^0 \to \mathbb{R}$ by

$$\eta_N(\mu) \overset{\triangle}{=} \mu^0 \psi_N(x)_+ + \sum_{j \in \mathbf{q}} \mu^j [\psi_N(x)_+ - f_N^j(x)] + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f_N^j(x) \Big\|^2. \tag{16}$$

Since Σ_q^0 is compact, it follows that $\sup_{\mu \in \Sigma_q^0} |\eta_N(\mu) - \eta(\mu)| \to 0$, as $N \to \infty$, almost surely. Hence, for any $\epsilon > 0$, there exists an N_0 such that for all $N \geq N_0$ and $\mu \in \Sigma_q^0$, $|\eta_N(\mu) - \eta(\mu)| \leq \epsilon$, almost surely. Since $\theta(x) = -\min_{\mu \in \Sigma_q^0} \eta(\mu)$ and $\theta_N(x) = -\min_{\mu \in \Sigma_q^0} \eta_N(\mu)$, it follows that $|\theta_N(x) - \theta(x)| \leq \epsilon$ for all $N \geq N_0$ almost surely. This completes the proof. □

3.2 Asymptotic Distribution of Estimator

We next examine the asymptotic distribution of an appropriately shifted and scaled $\theta_N(x)$ for a given $x \in \mathbb{R}^n$. Before we state the main result of this section (Theorem 2), we need to establish some notation.

Let for any $x \in \mathbb{R}^n$,

$$\hat{\Sigma}_q^0(x) \overset{\triangle}{=} \Big\{\mu \in \Sigma_q^0 \;\Big|\; \theta(x) = \mu^0\psi(x)_+ + \sum_{j\in\mathbf{q}} \mu^j[\psi(x)_+ - f^j(x)] + \tfrac{1}{2}\Big\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j(x)\Big\|^2\Big\}, \qquad (17)$$

$\hat{\mathbf{q}}(x) \overset{\triangle}{=} \{j \in \mathbf{q} \mid \psi(x) = f^j(x)\}$, and

$$\hat{\mathbf{q}}(x)_+ \overset{\triangle}{=} \begin{cases} \hat{\mathbf{q}}(x) \cup \{0\} & \text{if } \psi(x) = 0 \\ \hat{\mathbf{q}}(x) & \text{if } \psi(x) > 0 \\ \{0\} & \text{otherwise.} \end{cases} \qquad (18)$$

We use v' to denote the transpose of a vector v and define the following quantities:

$$f(x) \overset{\triangle}{=} (f^1(x), f^2(x), ..., f^q(x))', \qquad (19)$$

$$f_N(x) \overset{\triangle}{=} (f_N^1(x), f_N^2(x), ..., f_N^q(x))', \qquad (20)$$

$$\nabla \overline{f}(x) \overset{\triangle}{=} (\nabla f^0(x)', \nabla f^1(x)', ..., \nabla f^q(x)')', \qquad (21)$$

and

$$\nabla \overline{f}_N(x) \overset{\triangle}{=} (\nabla f_N^0(x)', \nabla f_N^1(x)', ..., \nabla f_N^q(x)')'. \qquad (22)$$

We need the following light-tail assumption to ensure a central limit theorem.

Assumption 2 *We assume that for a given $x \in \mathbb{R}^n$, $E[F^j(x,\omega)^2] < \infty$ for all $j \in \mathbf{q}$ and $E[(\partial F^j(x,\omega)/\partial x^i)^2] < \infty$ for all $j \in \mathbf{q}_0$ and $i = 1, 2, ..., n$.* □

For any $x \in \mathbb{R}^n$, we let $\overline{Y}(x)$ denote the $q + (q+1)n$-dimensional normal random vector with zero mean and variance-covariance matrix $\overline{V}(x)$, where $\overline{V}(x)$ is the variance-covariance matrix of the random vector $(F^1(x,\omega), F^2(x,\omega), ..., F^q(x,\omega), \nabla_x F^0(x,\omega)', \nabla_x F^1(x,\omega)', ..., \nabla_x F^q(x,\omega)')'$. Moreover, we define the q-dimensional random vector $Y_{-1}(x)$ and the n-dimensional random vectors $Y_j(x)$, $j \in \mathbf{q}_0$, such that $\overline{Y}(x) = (Y_{-1}(x)', Y_0(x)', Y_1(x)', ..., Y_q(x)')'$.

We use $\Rightarrow$ to denote convergence in distribution. The following vector-valued central limit theorem is well known; see, for example, Theorem 29.5 in [8].

Proposition 4 *Suppose that Assumption 2 holds at $x \in \mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$. Then,*

$$N^{1/2}\left(\begin{pmatrix} f_N(x) \\ \nabla \overline{f}_N(x) \end{pmatrix} - \begin{pmatrix} f(x) \\ \nabla \overline{f}(x) \end{pmatrix}\right) \Rightarrow \overline{Y}(x) \qquad (23)$$

as $N \to \infty$. □

We next provide the asymptotic distribution of a scaled and shifted $\theta_N(x)$.

Theorem 2 *Suppose that Assumption 2 holds at $x \in \mathbb{R}^n$ and that Assumption 1 is satisfied on an open set containing $x \in \mathbb{R}^n$. Then,*

$$N^{1/2}(\theta_N(x)-\theta(x)) \Rightarrow -\min_{\mu\in\hat{\Sigma}_q^0(x)} \Big\{\mu^0 W(x)+\sum_{j\in\mathbf{q}}\mu^j[W(x)-Y_{-1}^j(x)]+\sum_{j\in\mathbf{q}_0}\mu^j\Big\langle \sum_{k\in\mathbf{q}_0}\mu^k\nabla f^k(x), Y_j(x)\Big\rangle\Big\} \tag{24}$$

as $N\to\infty$, where $W(x) \triangleq \max_{j\in\hat{\mathbf{q}}(x)_+} Y_{-1}^j(x)$, with $Y_{-1}^0(x)\triangleq 0$.

Proof: The proof is based on the Delta Theorem 7.59 in [34]. Let $g:\mathbb{R}^{q+(q+1)n}\to\mathbb{R}$ be defined for any $\overline{\zeta}=(\zeta_{-1},\zeta_0',\zeta_1',...,\zeta_q')\in\mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1}\in\mathbb{R}^q$, $\zeta_j\in\mathbb{R}^n$, $j\in\mathbf{q}_0$, by

$$g(\overline{\zeta})\triangleq -\min_{\mu\in\Sigma_q^0}\Big\{\mu^0 w(\overline{\zeta})+\sum_{j\in\mathbf{q}}\mu^j[w(\overline{\zeta})-\zeta_{-1}^j]+\tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\zeta_j\Big\|^2\Big\}, \tag{25}$$

where $w:\mathbb{R}^{q+(q+1)n}\to\mathbb{R}$ is defined by $w(\overline{\zeta})\triangleq\max\{0,\max_{j\in\mathbf{q}}\zeta_{-1}^j\}$. Since $\sum_{j\in\mathbf{q}_0}\mu^j=1$ for all $\mu\in\Sigma_q^0$, it follows that

$$g(\overline{\zeta})=-w(\overline{\zeta})-\phi(\overline{\zeta}), \tag{26}$$

where $\phi:\mathbb{R}^{q+(q+1)n}\to\mathbb{R}$ is defined by

$$\phi(\overline{\zeta})\triangleq\min_{\mu\in\Sigma_q^0}\Big\{-\sum_{j\in\mathbf{q}}\mu^j\zeta_{-1}^j+\tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\zeta_j\Big\|^2\Big\}. \tag{27}$$

Let

$$\hat{\mathbf{q}}_w(\overline{\zeta})\triangleq\{j\in\mathbf{q}\mid \max_{k\in\mathbf{q}}\zeta_{-1}^k=\zeta_{-1}^j\}, \tag{28}$$

and

$$\hat{\mathbf{q}}_w(\overline{\zeta})_+\triangleq\begin{cases}\hat{\mathbf{q}}_w(\overline{\zeta})\cup\{0\} & \text{if } w(\overline{\zeta})=0\\ \hat{\mathbf{q}}_w(\overline{\zeta}) & \text{if } w(\overline{\zeta})>0\\ \{0\} & \text{otherwise.}\end{cases} \tag{29}$$

Moreover, let

$$\hat{\Sigma}_\phi(\overline{\zeta})\triangleq\Big\{\mu\in\Sigma_q^0\;\Big|\;\phi(\overline{\zeta})=-\sum_{j\in\mathbf{q}}\mu^j\zeta_{-1}^j+\tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\zeta_j\Big\|^2\Big\}. \tag{30}$$

It follows from Danskin Theorem; see, for example, Theorem 7.21 in [34], that $w(\cdot)$ and $\phi(\cdot)$ are locally Lipschitz continuous and directional differentiable with directional derivatives at $\overline{\zeta}\in\mathbb{R}^{q+(q+1)n}$ in the direction $\overline{\xi}\in\mathbb{R}^{q+(q+1)n}$ given by

$$dw(\overline{\zeta};\overline{\xi})=\max_{j\in\hat{\mathbf{q}}_w(\overline{\zeta})_+}\xi_{-1}^j, \tag{31}$$

with $\xi_{-1}^0\triangleq 0$, and

$$d\phi(\overline{\zeta};\overline{\xi})=\min_{\mu\in\hat{\Sigma}_\phi(\overline{\zeta})}\Big\{-\sum_{j\in\mathbf{q}}\mu^j\xi_{-1}^j+\sum_{j\in\mathbf{q}_0}\mu^j\Big\langle\sum_{k\in\mathbf{q}_0}\mu^k\zeta_k,\xi_j\Big\rangle\Big\}. \tag{32}$$

Consequently, $g(\cdot)$ is locally Lipschitz continuous and directional differentiable with directional derivatives at $\overline{\zeta} \in \mathbb{R}^{q+(q+1)n}$ in the direction $\overline{\xi} \in \mathbb{R}^{q+(q+1)n}$ given by

$$dg(\overline{\zeta};\overline{\xi}) = -\max_{j\in\hat{\mathbf{q}}_w(\overline{\zeta})_+} \xi^j_{-1} - \min_{\mu\in\hat{\Sigma}_\phi(\overline{\zeta})} \Big\{ -\sum_{j\in\mathbf{q}} \mu^j \xi^j_{-1} + \sum_{j\in\mathbf{q}_0} \mu^j \Big\langle \sum_{k\in\mathbf{q}_0} \mu^k \zeta_k, \xi_j \Big\rangle \Big\}. \tag{33}$$

Hence, it follows from Proposition 7.57 in [34] that $g(\cdot)$ is Hadamard directional differentiable.

In view of Proposition 4, Delta Theorem 7.59 in [34] gives that

$$N^{1/2}(g((f_N(x), \nabla\overline{f}_N(x)')') - g((f(x), \nabla\overline{f}(x)')')) \Rightarrow dg((f(x), \nabla\overline{f}(x)')'; \overline{Y}(x)). \tag{34}$$

The result now follows from the facts that $g((f_N(x), \nabla\overline{f}_N(x)')') = \theta_N(x)$, $g((f(x), \nabla\overline{f}(x)')') = \theta(x)$, $\hat{\mathbf{q}}_w((f(x), \nabla\overline{f}(x)')')_+ = \hat{\mathbf{q}}(x)_+$, and $\hat{\Sigma}_\phi((f(x), \nabla\overline{f}(x)')') = \hat{\Sigma}^0_q(x)$ and from rearranging terms. □

In general, the right-hand side in (24) is not a normal random variable. Hence, $\theta_N(x)$ cannot be expected to be approximately normal even for large N. In special cases, we find the following interesting corollaries.

Corollary 1 *Suppose that Assumption 2 holds at $x \in \mathbb{R}^n$ and that Assumption 1 is satisfied on an open set containing $x \in \mathbb{R}^n$. Then, the following statements hold:*

(i) *If the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_0$, are linearly independent, then $\hat{\Sigma}^0_q(x) = \{\hat{\mu}(x)\}$ is a singleton and*

$$\begin{aligned} &N^{1/2}(\theta_N(x) - \theta(x)) \\ \Rightarrow\ &-\hat{\mu}^0(x)W(x) - \sum_{j\in\mathbf{q}} \hat{\mu}^j(x)[W(x) - Y^j_{-1}(x)] - \sum_{j\in\mathbf{q}_0} \hat{\mu}^j(x) \Big\langle \sum_{k\in\mathbf{q}_0} \hat{\mu}^k \nabla f^k(x), Y_j(x) \Big\rangle, \end{aligned} \tag{35}$$

as $N \to \infty$.

(ii) *If x is a local minimizer of P and the vectors $\nabla f^j(x)$, $j \in \hat{\mathbf{q}}(x)$, are linearly independent, then $\hat{\Sigma}^0_q(x) = \{\hat{\mu}(x)\}$ is a singleton and*

$$N^{1/2}\theta_N(x) \Rightarrow -W(x) + \sum_{j\in\hat{\mathbf{q}}(x)_+} \hat{\mu}^j(x) Y^j_{-1}(x) \tag{36}$$

as $N \to \infty$. Moreover, if in addition $\hat{\mathbf{q}}(x) = \{j(x)\}$ is a singleton, then

$$N^{1/2}\theta_N(x) \Rightarrow \begin{cases} -\max\{0, Y^{j(x)}_{-1}\} + \hat{\mu}^{j(x)}(x) Y^{j(x)}_{-1}(x) & \text{if } f^{j(x)}(x) = 0 \\ 0 & \text{if } f^{j(x)}(x) < 0 \end{cases} \tag{37}$$

as $N \to \infty$.

Proof: If the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_0$, are linearly independent, then the matrix $A(x) = (\nabla f^0(x), \nabla f^1(x), ..., \nabla f^q(x))$ has rank $q+1$. Hence, $A(x)'A(x)$ is positive definite and the objective function in (8) is strictly convex. Consequently, $\hat{\Sigma}^0_q(x)$ is a singleton and part (i) follows directly.

Next, consider part (ii). Since $x \in \mathbb{R}^n$ is a local minimizer of P, $\psi(x) \leq 0$ and, from Proposition 2, $\theta(x) = 0$. Hence, it follows from (8) that there exists a $\hat{\mu}(x) \in \hat{\Sigma}^0_q(x)$ such that

$\sum_{j\in\mathbf{q}_0}\hat\mu^j(x)\nabla f^j(x)=0$ and $\sum_{j\in\mathbf{q}}\hat\mu^j(x)[\psi(x)_+-f^j(x)]=0$. Consequently, $\hat\mu^j(x)=0$ for all $j\in\mathbf{q}$ such that $j\notin\hat{\mathbf{q}}(x)_+$. We deduce from the Karush-Kuhn-Tucker conditions for P (see for example Theorem 3.3.1 in [7]) that under the stated linear independence assumption, $\hat\Sigma_q^0(x)$ is a singleton. Since $Y_{-1}^0(x)=0$ by definition, (24) reduces to (36). Finally, (37) follows from (36). □

Corollary 2 *Suppose that Assumption 2 holds at $x\in\mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x\in\mathbb{R}^n$. If all constraints are deterministic, i.e., $F^j(\cdot,\omega)=F^j(x)$, $j\in\mathbf{q}$, then*

$$N^{1/2}(\theta_N(x)-\theta(x))\Rightarrow-\min_{\mu\in\hat\Sigma_q^0(x)}\mu^0\Big\langle\sum_{k\in\mathbf{q}_0}\mu^k\nabla f^k(x),Y_0(x)\Big\rangle,\qquad(38)$$

as $N\to\infty$.

Proof: This result follows by similar argument as those leading to Theorem 2. □

We see from (38) that $\theta_N(x)$ is approximately normal when $\hat\Sigma_q^0(x)$ is a singleton. Moreover, the limiting distribution degenerates to the constant zero when $\theta(x)=0$.

The next corollary deals with the special case of no constraints.

Corollary 3 *Suppose that Assumption 2 holds at $x\in\mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x\in\mathbb{R}^n$. If there are no constraints in P, then*

$$N^{1/2}(\theta_N(x)-\theta(x))\Rightarrow\mathcal{N}(0,\nabla f^0(x)'V_0(x)\nabla f^0(x)),\qquad(39)$$

as $N\to\infty$, where $V_0(x)$ is the n-by-n variance-covariance matrix of $Y_0(x)$ (and $\nabla_xF^0(x,\omega)$) and $\mathcal{N}(0,\sigma^2)$ denotes a zero-mean normal random variable with variance σ^2.

Proof: This result follows from Theorem 2. It can also be shown using Delta Theorem 7.59 in [34] and the fact (see p. 6 in [25]) that in this case we obtain the simplifications

$$\theta(x)=-\tfrac{1}{2}\|\nabla f(x)\|^2\qquad(40)$$

and

$$\theta_N(x)=-\tfrac{1}{2}\|\nabla f_N(x)\|^2.\qquad(41)$$

□.

We next consider the bias $\overline{E}\theta_N(x)-\theta(x)$, where $\overline{E}$ denotes the expectation with respect to $\overline{\mathcal{P}}$. Convergence in distribution do not necessarily imply convergence of expectations. Under an uniform integrability property, however, the convergence of expectations is ensured; see for example p. 338 of [8]. The property holds under several assumptions, one of which is used in the next result.

Proposition 5 *Suppose that Assumption 2 holds at $x\in\mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x\in\mathbb{R}^n$. Moreover, suppose that there exists an $\epsilon>0$ such that*

$$\sup_{N\in\mathbb{N}}\overline{E}[|N^{1/2}(\theta_N(x)-\theta(x))|^{1+\epsilon}]<\infty.\qquad(42)$$

Then,

$$
\begin{aligned}
&\overline{E}\theta_N(x) - \theta(x) && (43)\\
= \;& N^{-1/2}\overline{E}\Big[-\min_{\mu\in\hat{\Sigma}_q^0(x)} \Big\{\mu^0 W(x) + \sum_{j\in\mathbf{q}} \mu^j [W(x) - Y_{-1}^j(x)] + \sum_{j\in\mathbf{q}_0} \mu^j \Big\langle \sum_{k\in\mathbf{q}_0} \mu^k \nabla f^k(x), Y_j(x)\Big\rangle\Big\}\Big]\\
+ \;& o(N^{-1/2}).
\end{aligned}
$$

Moreover, if $\hat{\Sigma}_q^0(x)$ *is a singleton, then*

$$\overline{E}\theta_N(x) - \theta(x) = -N^{-1/2}\overline{E}[W(x)] + o(N^{-1/2}). \tag{44}$$

Proof: From Theorem 25.12 in [8] and Theorem 2, we directly obtain (43). Since Y_{-1}^j, $j \in \mathbf{q}$, and $Y_j(x)$, $j \in \mathbf{q}_0$, have zero mean and $\sum_{j\in\mathbf{q}_0} \mu^j = 1$ for all $\mu \in \Sigma_q^0$, (44) also holds. □

Conditions that ensure that $\hat{\Sigma}_q^0(x)$ is a singleton is given in Corollary 1.

We observe that the bias identified above is similar to the well-known bias of the optimal value of $\min_{x\in X_\psi} f_N^0(x)$ relative to the optimal value of $\min_{x\in X_\psi} f^0(x)$; see, for example p. 167 in [34]. In that case, the bias is always nonpositive. In the present case, $\overline{E}\theta_N(x)$ may be larger than $\theta(x)$. However, in the absence of constraints in P, it follows directly from (40) and (41) and Jensen's inequality that for any $N \in \mathbb{N}$,

$$\overline{E}\theta_N(x) \leq \theta(x). \tag{45}$$

4 Validation Analysis

In this section, we develop procedures for assessing the quality of a candidate point $x \in \mathbb{R}^n$. Specifically, we develop confidence intervals and probabilistic bounds on $\theta(x)$ and $\psi(x)$. Using such bounds, we may claim with some confidence that x satisfies the conditions $\psi(x) \leq \delta$ and $\theta(x) \geq -\epsilon$ for a given $\delta \geq 0$ and $\epsilon > 0$. We first consider the situation with no constraints in P, second deal with near feasibility, and third bound the optimality function of the full problem.

4.1 Unconstrained Optimization

Suppose that there are no constraints in P and let $x \in \mathbb{R}^n$ be a candidate solution. In view of Corollary 3, $\theta_N(x)$ is approximately normal with mean $\theta(x)$ and variance $\nabla f(x)' V_0(x) \nabla f(x)/N$ for large N. Hence, it is straightforward to construct a confidence interval for $\theta(x)$. Let

$$V_{0,N}(x) \overset{\triangle}{=} \frac{1}{N-1}\sum_{i=1}^{N} (\nabla_x F(x,\omega_i) - f_N(x))'(\nabla_x F(x,\omega_i) - f_N(x)). \tag{46}$$

be the standard unbiased estimator of V_0. Then for large N,

$$\Big(\theta_N(x) - z_\alpha\sqrt{\nabla f_N(x)' V_{0,N}(x) \nabla f_N(x)/N},\ 0\Big] \tag{47}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\theta(x)$, where z_α is the standard normal α-quantile. In (47) and other confidence intervals below we use a quantile of the standard normal distribution instead of one of the t-distribution as the sample size is typically relatively large.

We observe that the approximate normality of $\theta_N(x)$ does not directly reflect the fact that $\theta_N(x) \leq 0$ almost surely. However, in practice, validation analysis is almost always carried out at an $x \in \mathbb{R}^n$ with $\theta(x) < 0$ in which case the truncation at zero is insignificant for large N. Our numerical experiments indicate that the normal model of $\theta_N(x)$ is quite accurate for both $\theta(x) < 0$ and $\theta(x) = 0$; see Section 6. The confidence interval (47) is one-sized, as are the confidence intervals derived below. While it is easy to convert (47) into a two-sided confidence interval, we believe that one-sided confidence intervals are more suitable in the present context as $\theta(x) \geq -\epsilon$ is a natural (though conceptual) criterion for stopping an algorithm applied to P. Hence, if (47) is contained in $(-\epsilon, 0]$, then we would be $100(1-\alpha)\%$ confident that $\theta(x) \geq -\epsilon$ is satisfied.

4.2 Near Feasibility in P

We next consider the full problem P and develop a procedure for determining whether $x \in \mathbb{R}^n$ is nearly feasible, i.e., $\psi(x) \leq \delta$ for some $\delta \geq 0$. We adopt a simple batching approach to estimate the value of $\psi(x)$. In the ranking and selection literature we find more sophisticated and potentially more efficient ways of determining whether x is nearly feasible; see for example [18] and references therein. It is also possible to estimate $f^j(x)$ independently for each constraints $j \in \mathbf{q}$; see [30]. However, we do not explore those possibilities further.

Since the function $m : \mathbb{R}^q \to \mathbb{R}$ defined for any $y \in \mathbb{R}^q$ by $m(y) \overset{\triangle}{=} \max_{j\in\mathbf{q}} y^j$ is convex, it follows by Jensen's inequality that

$$\psi(x) \leq \overline{E}\psi_N(x). \tag{48}$$

We construct a confidence interval for $\overline{E}\psi_N(x)$, which, in view of (48), provides a conservative confidence interval for $\psi(x)$.

For given N and M, let $\psi_{N,k}(x)$, $k = 1, 2, ..., M$, be independent random variables distributed as $\psi_N(x)$. Then,

$$\overline{\psi}_{N,M}(x) \overset{\triangle}{=} \frac{1}{M}\sum_{k=1}^{M}\psi_{N,k}(x) \tag{49}$$

is an unbiased estimator of $\overline{E}\psi_N(x)$. If $E[F^j(x,\omega)^2] < \infty$ for all $j \in \mathbf{q}$, then a central limit theorem holds for $\overline{\psi}_{N,M}(x)$, i.e., $\overline{\psi}_{N,M}(x)$ is approximately normal with mean $\overline{E}\psi_N(x)$ and variance $Var[\psi_N(x)]/M$ for large M. Let $s^2_{\psi,N,M}(x)$ be the standard unbiased estimator of $Var[\psi_N(x)]$ given by

$$s^2_{\psi,N,M}(x) = \frac{1}{M-1}\sum_{k=1}^{M}(\psi_{N,k}(x) - \overline{\psi}_{N,M}(x))^2. \tag{50}$$

Then, it follows that

$$(-\infty, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}) \tag{51}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\overline{E}\psi_N(x)$ for large M. In view of (48), it follows that (51) is a conservative $100(1-\alpha)\%$-confidence interval for $\psi(x)$ for large M. This confidence interval can be used to assess whether the candidate solution x satisfies $\psi(x) \leq \delta$.

4.3 Constrained Optimization

We propose two approaches for obtaining confidence intervals for $\theta(x)$. The first approach makes use of the following result.

Proposition 6 *Consider $x \in \mathbb{R}^n$ and suppose that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$. Then, for any $\mu \in \Sigma_q^0$,*

$$\theta(x) \geq \overline{E}\Big[-\mu^0\psi_N(x)_+ - \sum_{j\in\mathbf{q}}\mu^j(\psi_N(x)_+ + f_N^j(x)) - \tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\nabla f_N^j(x)\Big\|^2\Big]. \tag{52}$$

Proof: For any $\mu \in \Sigma_q^0$, let $\tilde{\eta} : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ be defined by

$$\tilde{\eta}(\overline{\zeta}) \overset{\triangle}{=} \max\{0, \max_{j\in\mathbf{q}}\zeta_{-1}^j\} - \sum_{j\in\mathbf{q}}\mu^j\zeta_{-1}^j + \tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\zeta_j\Big\|^2 \tag{53}$$

for any $\overline{\zeta} = (\zeta_{-1}', \zeta_0', \zeta_1', ...\zeta_q') \in \mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1} \in \mathbb{R}^q$ and $\zeta_j \in \mathbb{R}^n$, $j \in \mathbf{q}_0$. Since $\tilde{\eta}(\cdot)$ is convex, it follows from Jensen's inequality that

$$\overline{E}\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')') \geq \tilde{\eta}((f(x)', \nabla\overline{f}(x)')'). \tag{54}$$

From (8) and (54), we see that

$$\begin{aligned}\tilde{\eta}((f(x)', \nabla\overline{f}(x)')') &= \mu^0\psi(x)_+ + \sum_{j\in\mathbf{q}}\mu^j(\psi(x)_+ + f^j(x)) + \tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\nabla f^j(x)\Big\|^2 && (55)\\ &\geq -\theta(x). && (56)\end{aligned}$$

The result then follows from the fact that $\overline{E}\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')$ equals the negative of the right-hand side in (52). □

In view of Proposition 6, we construct a conservative confidence interval for $\theta(x)$ by computing a confidence interval for the right-hand side in (52). We adopt a batching approach and, for given N and M, let $\eta_{N,k}$, $k = 1, 2, ..., M$, be independent random variables distributed as $\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')$. Then,

$$\overline{\eta}_{N,M} \overset{\triangle}{=} \frac{1}{M}\sum_{k=1}^{M}\eta_{N,k} \tag{57}$$

is an unbiased estimator of $\overline{E}[\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')]$. Under sufficient integrability assumptions for $(f_N(x)', \nabla\overline{f}_N(x)')$, a central limit theorem holds for $\overline{\eta}_{N,M}$ and, consequently, $\overline{\eta}_{N,M}$ is approximately normal with mean $\overline{E}[\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')]$ and variance $Var[\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')]/M$ for large M. Let $s_{\eta,N,M}^2$ be the standard unbiased estimator of $Var[\tilde{\eta}((f_N(x)', \nabla\overline{f}_N(x)')')]$ given by

$$s_{\eta,N,M}^2 = \frac{1}{M-1}\sum_{k=1}^{M}(\eta_{N,k} - \overline{\eta}_{N,M})^2. \tag{58}$$

Then, it follows that

$$(-\overline{\eta}_{N,M} - z_\alpha s_{\eta,N,M}(x)/\sqrt{M}, 0] \tag{59}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\overline{E}[-\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ for large M. In view of (52), it follows that (59) is a conservative $100(1-\alpha)\%$-confidence interval for $\theta(x)$ for large M. To compute the above confidence interval, it is necessary to select a $\mu \in \Sigma_q^0$. In view of the proof of Proposition 6, we see that a tighter confidence interval can be expected when $\mu \in \hat{\Sigma}_q^0(x)$. Hence, we recommend to select μ as the optimal solution of (14) for some large N. We note, however, that even when using $\mu \in \hat{\Sigma}_q^0(x)$, the inequality in (52) may be strict.

The second approach to constructing a confidence interval for $\theta(x)$ is motivated by a procedure for obtaining bounds on the optimal value of optimization problems with chance constraints; see Section 5.7.2 in [34]. The approach is somewhat more limited than the first approach as it requires the following independence assumption.

Assumption 3 *We assume that for a given $x \in \mathbb{R}^n$, the random vectors $(f^j(x), \nabla f^j(x)')'$, $j \in \mathbf{q}$, are statistically independent. Moreover, we assume that $\nabla f^0(x)$ is statistically independent of $(f^j(x), \nabla f^j(x)')'$ for all $j \in \mathbf{q}$.* □

Assumption 3 trivially holds for all $x \in \mathbb{R}^n$ in the important special case when all but one of the random functions $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, are deterministic.

It is beneficial to "decompose" the optimality function into feasibility and optimality parts. From (7) we see that $\theta(x) = -\psi(x)_+ + u(x)$, where

$$\begin{aligned} u(x) \quad &\overset{\triangle}{=} \quad \min_{h \in \mathbb{R}^n, z \in \mathbb{R}} z + \tfrac{1}{2}\|h\|^2 \\ &s.t. \quad \langle \nabla f^0(x), h \rangle \leq z \\ &\qquad\quad f^j(x) + \langle \nabla f^j(x), h \rangle \leq z, j \in \mathbf{q}. \end{aligned} \tag{60}$$

Here, $-\psi(x)_+$ is a measure of feasibility and $u(x)$ is a measure of optimality. Similarly, let

$$\begin{aligned} u_N(x) \quad &\overset{\triangle}{=} \quad \min_{h \in \mathbb{R}^n, z \in \mathbb{R}} z + \tfrac{1}{2}\|h\|^2 \\ &s.t. \quad \langle \nabla f_N^0(x), h \rangle \leq z \\ &\qquad\quad f_N^j(x) + \langle \nabla f_N^j(x), h \rangle \leq z, j \in \mathbf{q}. \end{aligned} \tag{61}$$

The next lemma provides a useful relationship between $u(x)$ and $u_N(x)$.

Lemma 1 *Suppose that Assumptions 2 and 3 hold at $x \in \mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$. Then,*

$$\liminf_{N\to\infty} \overline{\mathcal{P}}[u_N(x) \leq u(x)] \geq \frac{1}{2^{q+1}}. \tag{62}$$

Proof: Suppose that $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is a feasible point in (60). We want to determine the probability, denoted $\hat{p}_N$, that $(\hat{h}, \hat{z})$ is feasible in (61). Since $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is feasible for (60), we obtain that

$$\begin{aligned} \hat{p}_N &\triangleq \overline{\mathcal{P}}\Big[\Big\{\langle \nabla f_N^0(x), \hat{h}\rangle \leq \hat{z}\Big\} \bigcap \Big(\bigcap_{j\in\mathbf{q}} \Big\{f_N^j(x) + \langle \nabla f_N^j(x), \hat{h}\rangle \leq \hat{z}\Big\}\Big)\Big] \\ &\geq \overline{\mathcal{P}}\Big[\Big\{\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \leq 0\Big\} \bigcap \\ &\quad \Big(\bigcap_{j\in\mathbf{q}} \Big\{f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \leq 0\Big\}\Big)\Big]. \end{aligned} \tag{63}$$

Under Assumption 3, it follows that

$$\hat{p}_N \geq \overline{\mathcal{P}}\Big[\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \leq 0\Big] \prod_{j\in\mathbf{q}} \overline{\mathcal{P}}\Big[f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \leq 0\Big]. \tag{64}$$

In view of Proposition 4, it follows that $N^{1/2}\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle$ converges in distribution to a zero-mean normal random variable. Hence,

$$\lim_{N\to\infty} \overline{\mathcal{P}}\Big[\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \leq 0\Big] \geq 1/2. \tag{65}$$

We observe that the limit in (65) is not equal to 1/2 as the zero-mean normal random variable may have zero variance. Similarly, for all $j \in \mathbf{q}$, $N^{1/2}(f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle)$ converges in distribution to a zero-mean normal random variable. Hence, for all $j \in \mathbf{q}$,

$$\lim_{N\to\infty} \overline{\mathcal{P}}\Big[f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \leq 0\Big] \geq 1/2. \tag{66}$$

Consequently, $\liminf_{N\to\infty} \hat{p}_N \geq 1/2^{q+1}$. Since this result holds for any $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ that is feasible in (60), it also holds for the optimal solution in (60). If $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is the optimal solution in (60) and it is also feasible in (61), then

$$u_N(x) \leq \hat{z} + \tfrac{1}{2}\|\hat{h}\|^2 = u(x). \tag{67}$$

This completes the proof. □

Lemma 1 provides the basis for the following procedure for obtaining a probabilistic lower bound on $u(x)$. This procedure is essentially identical to the one proposed in Section 5.7.2 of [34] in the context of chance constraints.

Let $u_{N,k}(x)$, $k = 1, 2, ..., K$, be independent random variables distributed as $u_N(x)$. After obtaining realizations of these random variables, we order them with respect to their values. Let $\tilde{u}_{N,1}, \tilde{u}_{N,2}, ..., \tilde{u}_{N,K}$, with $\tilde{u}_{N,k} \leq \tilde{u}_{N,k+1}$, be this ordered sequence. That is, $\tilde{u}_{N,1}$ is the smallest value of $u_{N,k}(x)$, $k = 1, 2, ..., K$, $\tilde{u}_{N,2}$ is the second smallest, etc. Suppose that $\hat{\gamma}_N$ is a lower bound on $\overline{\mathcal{P}}[u_N(x) \leq u(x)]$ and suppose that for a given $\beta \in (0, 1)$, K and L satisfy

$$\sum_{k=0}^{L-1} \begin{pmatrix} K \\ k \end{pmatrix} \hat{\gamma}_N^k (1 - \hat{\gamma}_N)^{K-k} \leq \beta. \tag{68}$$

Then, using the same arguments as in 5.7.2 of [34], we obtain that

$$\overline{\mathcal{P}}[\tilde{u}_{N,L} > u(x)] \leq \beta. \tag{69}$$

Hence, $(\tilde{u}_{N,L}, 0]$ is a $100(1-\beta)\%$-confidence interval for $u(x)$. In view of Lemma 1 and its proof, we recommend a number slightly smaller than $1/2^{q+1}$ as an estimate of the lower bound $\hat{\gamma}_N$ when N is moderately large.

Suppose that the confidence interval for $\psi(x)$ in (51) is computed independently of the confidence interval for $u(x)$. Then,

$$\Big(- \max\{0, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}\} + \tilde{u}_{N,L},\ 0\Big] \tag{70}$$

is an approximate $100(1-\alpha)(1-\beta)\%$-confidence interval for $\theta(x)$ for large M and N. We observe that the first approach to computing a confidence interval for $\theta(x)$ requires the solution of only one convex quadratic optimization problem to obtain $\mu \in \Sigma_q^0$. The second approach requires K such solutions. If $L = 1$, then $K \geq \log\beta/\log(1-\hat{\gamma}_N)$. Hence, K is typically moderate. For example, if $\beta = 0.01$ and $\hat{\gamma}_N = 0.49$, then $K = 7$ suffices.

5 Algorithm and Consistent Approximations

There are numerous algorithms for solving stochastic programs similar to P including decomposition algorithms in cases with special structure (see, e.g., [16, 13]), stochastic approximations (see, e.g., [10, 6, 20, 23]), other versions of stochastic search (see, e.g., [36]), and sample average approximations (see, e.g., [34]). A detail review of these algorithms is beyond the scope the paper. However, we note that special structure may not be present and stochastic approximations may be problematic to apply to P as that problem possibly involves constraints given by nonconvex expect value functions. In this section, we use sample average approximations, the optimality function $\theta(\cdot)$, and $\theta_N(x)$ to construct an algorithm for P that converges almost surely under a smoothness assumption.

5.1 Consistent Approximations

We adopt the framework of sample average approximation and define the sample average problem

$$P_N: \quad \min_{x \in \mathbb{R}^n} \{f_N^0(x) \mid f_N^j(x) \leq 0, j \in \mathbf{q}\}. \tag{71}$$

It is well-known that under suitable assumptions, the optimal value and the set of optimal solutions of P_N converges in some sense to the optimal value and the set of optimal solutions of P, respectively; see for example Chapter 5 of [34]. While such results provide guidance to the selection of N, they do not directly translate into an implementable algorithm for solving P. In particular, if P_N is nonconvex, then a globally optimal solution of P_N may be beyond reach. In this section, we show that P_N and an associated optimality function are weakly consistent approximations (see Section 3.3 in [25] and description below) of P and $\theta(\cdot)$. This result directly leads to an implementable algorithm for P. Note that the concept of consistent approximations in [25] is not directly related to consistency of estimators.

We need the following strengthening of Assumption 1.

Assumption 4 *We assume that for a given set $S \subset \mathbb{R}^n$, the following hold for all $j \in \mathbf{q}_0$:*

(i) *For almost all $\omega \in \Omega$, $F^j(\cdot, \omega)$ is continuously differentiable on S.*

(ii) *There exists nonnegative valued measurable function such that $E[C(\omega)] < \infty$ and for every $x \in S$, $|F^j(x, \omega)| \leq C(\omega)$ and $\|\nabla_x F^j(x, \omega)\| \leq C(\omega)$ for almost all $\omega \in \Omega$.*

Assumption 4 is identical to those made in [2] and holds for example in the context of estimation of mixed logit models. Important models such as two-stage stochastic programs with recourse and conditional Value-at-Risk problems involve nonsmooth functions $F^j(\cdot, \omega)$ (see for example [17] and [27]) and hence do not satisfy Assumption 4 (i). However, recent efforts to apply smooth approximations of $F^j(\cdot, \omega)$ appear promising [1, 39] and facilitate the use of the algorithm below also in these nonsmooth cases.

If Assumption 4 holds on an open set S, then it follows from Theorems 7.44 and 7.48 in [34] that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are continuously differentiable on S and that

$$\nabla f^j(x) = E[\nabla_x F^j(x, \omega)], \tag{72}$$

for all $x \in S$ and $j \in \mathbf{q}_0$. Moreover, $f_N^j(\cdot, \overline{\omega})$, $j \in \mathbf{q}_0$, are continuously differentiable for almost every $\overline{\omega} \in \overline{\Omega}$. Hence, P_N is a smooth optimization problem almost surely. Consequently, the estimator $\theta_N(x)$ of $\theta(x)$ can be viewed as an optimality function $\theta_N : \mathbb{R}^n \to (\infty, 0]$ for P_N. It follows trivially that Proposition 2 holds with $\psi(\hat{x})$ and $\theta(\hat{x})$ replaced by $\psi_N(\hat{x})$ and $\theta_N(\hat{x})$, respectively, and $f^j(\hat{x})$ and $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_0$, replaced by $f_N^j(\hat{x})$ and $\nabla f_N^j(\hat{x})$, $j \in \mathbf{q}_0$, respectively, in (5) and (6). Similarly, Proposition 3 holds under suitable assumptions with $\theta(x)$, $f(x)$, and $f(\hat{x})$ replaced by $\theta_N(x)$, $f_N(x)$, and $f_N(\hat{x}_N)$, where $\hat{x}_N$ is the optimal solution of P_N. Hence, $\theta_N(x)$ can be viewed as a measure of the distance between x and a Fritz-John point of P_N.

We adopt the definition of weakly consistent approximations from Section 3.3 in [25]: The elements of the sequence $\{(P_N, \theta_N(\cdot)\}_{N=1}^{\infty}$ are weakly consistent approximations of $(P, \theta(\cdot))$ if (i) $P_N \to^{epi} P$, as $N \to \infty$, almost surely, and (ii) for any $x \in \mathbb{R}^n$ and sequence $\{x_N\}_{N=1}^{\infty} \subset \mathbb{R}^n$ with $x_N \to x$, as $N \to \infty$, $\limsup_{N \to \infty} \theta_N(x_N) \leq \theta(x)$, almost surely.

We proceed by showing that $\{(P_N, \theta_N(\cdot)\}_{N=1}^{\infty}$ indeed are weakly consistent approximations of $(P, \theta(\cdot))$. We first consider epiconvergence of P_N to P, which requires the following constraint qualification.

Assumption 5 *We assume for a given set $S \subset \mathbb{R}^n$ and for almost every $\overline{\omega} \in \overline{\Omega}$ that the following holds: For every $x \in S$ satisfying $\psi(x) \leq 0$, there exists a sequence $\{x_N\}_{N=1}^{\infty} \subset S$, with $\psi_N(x_N, \overline{\omega}) \leq 0$, such that $x_N \to x$, as $N \to \infty$.* □

When Assumptions 4 holds on a compact set S, then for all $j \in \mathbf{q}_0$, $f_N^j(x)$ converges to $f^j(x)$ uniformly on S almost surely; see for example Theorem 7.48 in [34]. It is well-known that epiconvergence follows under the stated assumptions. This result is given in the next proposition; see for example Theorem 3.3.2 in [25] for a proof.

Proposition 7 *Suppose that Assumptions 4 and 5 hold on an open and bounded set $S \subset \mathbb{R}^n$ and $X_\psi \subset S$. Then, $P_N \to^{epi} P$, as $N \to \infty$, almost surely.* □

We next consider the requirement on the relationship between $\theta_N(\cdot)$ and $\theta(\cdot)$, as $N \to \infty$. When Assumptions 4 holds on a compact set S, then for all $j \in \mathbf{q}_0$, $\nabla f_N^j(x)$ converges to $\nabla f^j(x)$ uniformly on S almost surely; see for example Theorem 7.48 in [34]. Hence, the next extension of Theorem 1 follows by essentially the same arguments as in that theorem's proof.

Proposition 8 *Suppose that Assumption 4 holds on an open set $S \subset \mathbb{R}^n$ and that $X \subset S$ is compact. Then, $\theta_N(x) \to \theta(x)$, as $N \to \infty$, uniformly on X, almost surely.* □

In view of Propositions 7 and 8, the next result follows directly from the definition of consistent approximations.

Theorem 3 *Suppose that Assumptions 4 and 5 hold on an open set $S \subset \mathbb{R}^n$, that $X \subset S$ is compact, and that $X_\psi \subset X$. Then, $\{(P_N, \theta_N(\cdot)\}_{N=1}^{\infty}$ are weakly consistent approximations of $(P, \theta(\cdot))$.* □

As we see in the next section, this result directly leads to an implementable algorithm for P.

5.2 Algorithm

We adapt Algorithm Model 3.3.14 in [25] to P. In essence, the resulting algorithm approximately solves the sequence of problems $\{P_N\}_{N \in \mathcal{K}}$, where $\mathcal{K}$ is an order set of strictly increasing positive integers with infinite cardinality. As N increases, the precision with which P_N is solved increases too. We measure the precision of a solution of P_N by means of the optimality function $\theta_N(\cdot)$. When a point of sufficient precision is obtained for P_N, then the algorithm starts solving $P_{N'}$, where N' is the next integer in $\mathcal{K}$ after N. We allow great flexibility in the choice of optimization algorithm for approximately solving $\{P_N\}_{N \in \mathcal{K}}$. Essentially, all convergent nonlinear programming solvers can be used.

For almost every $\overline{\omega} = (\omega_1, \omega_2, ...) \in \overline{\Omega}$ and any $N \in \mathbb{N}$, let $A_N : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be a deterministic algorithm map that defines one iteration of a nonlinear programming algorithm as applied to

P_N with the sample ω_1, ω_2, ..., ω_N. We assume that the algorithm map satisfies the following assumption.

Assumption 6 *For almost every $\overline{\omega} = (\omega_1, \omega_2, ...) \in \overline{\Omega}$ and any $N \in \mathbb{N}$, we assume that every accumulation point $\hat{x} \in \mathbb{R}^n$ of a sequence $\{x_i\}_{i=0}^{\infty}$ generated by the algorithm map $A_N(\cdot)$, i.e., $x_{i+1} \in A_N(x_i)$, satisfies $\theta_N(\hat{x}) = 0$ and $\psi_N(\hat{x}) \leq 0$.*

We next state the algorithm, where we use the notation $k(N)$ to denote the smallest $N' \in \mathcal{K}$ strictly greater than N.

Algorithm 1 (Solves P)

Input. Function $\Delta : \mathbb{N} \to (0, \infty)$ such that $\Delta(N) \to 0$, as $N \to \infty$; an ordered set $\mathcal{K}$ of strictly increasing positive integers with infinite cardinality; parameters $\epsilon, \delta > 0$; $N_0 \in \mathcal{K}$; $x_0 \in \mathbb{R}^n$; and realizations $\omega_1, \omega_2, ...$ obtained by independent sampling from $\mathcal{P}$.

Step 0. Set $i = 0$, $x_0^* = x_0$, and $N = N_0$.

Step 1. Using ω_1, ω_2, ..., ω_N, compute $x_{i+1} \in A_N(x_i)$.

Step 2. If $\theta_N(x_{i+1}) \geq -\epsilon\Delta(N)$ and $\psi_N(x_{i+1}) \leq \delta\Delta(N)$, then set $x_N^* = x_{i+1}$ and replace N by $k(N)$.

Step 3. Replace i by $i + 1$, and go to Step 1.

In Algorithm 1, P_N and $P_{N'}$ are not independent for any $N, N' \in \mathcal{K}$ as the sample is augmented and not regenerated. We note that the infinite sequence of realizations $\omega_1, \omega_2, ...$ can be generated as needed. That is, initially, generate $\omega_1, \omega_2, ..., \omega_{N_0}$, and then augment the sequence as N is increased.

The following convergence result for Algorithm 1 follows directly from Theorem 3.3.15 in [25].

Proposition 9 *Suppose that Assumptions 4, 5, and 6 hold on a sufficiently large open subset of $\mathbb{R}^n$. Moreover, suppose that Algorithm 1 has generated the sequences $\{x_N^*\}$ and $\{x_i\}_{i=0}^{\infty}$ and they are bounded. Then, every accumulation point $\hat{x}$ of $\{x_N^*\}$ satisfies $\theta(\hat{x}) = 0$ and $\psi(\hat{x}) \leq 0$ almost surely.* □

Algorithm 1 does not include a stopping criterion. One might run Algorithm 1 until a predetermined computing budget is exhausted and then carry out validation analysis on the candidate points $\{x_i\}$ or a subset thereof. For example, validation analysis may include computing confidence intervals for $\psi(x_i)$ and $\theta(x_i)$ using (51), (59), and/or (70). The confidence intervals need to be computed using a sample that is independent of the one used in Algorithm 1 to ensure the stated approximate coverage probabilities.

In practice, one might also want to carry out sequential validation analysis. That is, whenever a new x_N^* or x_i is generated, immediately assess its quality. If the quality is satisfactory, then

	Confidence intervals			
N	$\theta(\hat{x})$	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
10^5	$(-0.2897, 0]$	$(-0.6515, 0]$	$(-62.09, 0]$	$(-5774, 0]$
10^6	$(-0.0427, 0]$	$(-0.5771, 0]$	$(-57.98, 0]$	$(-5747, 0]$
10^7	$(-0.0043, 0]$	$(-0.4617, 0]$	$(-57.55, 0]$	$(-5743, 0]$
∞	0	-0.4420	-57.40	-5740

Table 1: Example 1: 95%-confidence intervals for $\theta(\hat{x})$ (column 2), $\theta(x_1)$ (column 3), $\theta(x_2)$ (column 4), and $\theta(x_3)$ (column 5) using (47) with varying sample size N. The last row gives the true values of $\theta(\hat{x})$, $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$.

stop the algorithm. Otherwise, let the algorithm continue. In this case, the last candidate solution generated by the algorithm is random. Hence, the confidence intervals derived in this paper may not have the stated approximate coverage probabilities when applied to that last solution; see [5] for a similar situation in the context of optimality gap estimation.

6 Numerical Examples

In this section, we present preliminary numerical tests of Algorithm 1 and the validation analysis procedures developed in Section 4 as applied to four simple examples. We recognize that variance reduction techniques (see for example Section 5.5 in [34]) may reduce computing times in validation analysis and optimization, but do not pursue that avenue in this paper. All calculations in this section are performed in Matlab 7.4 on a 2.16 GHz laptop computer with 1 GB of RAM and Windows XP.

6.1 Example 1: Validation Analysis for Unconstrained Problem

We consider an instance of P where there are no constraint, $n = 20$, and $F^0(\cdot,\cdot)$ is defined by

$$F^0(x,\omega) = \sum_{i=1}^{20} a^i (x^i - b^i \omega^i)^2 \tag{73}$$

where $a^i = i$, $b^i = 21 - i$, $i = 1, 2, ..., 20$, and $\omega = (\omega^1, \omega^2, ..., \omega^{20})'$ is a vector of independent and uniformly $[0, 1]$ distributed random variables. In this instance, both $\nabla f^0(x)$ and the unique global minimizer $\hat{x} = (10, 9.5, 9, 8.5, ..., 0.5)'$ are easy to compute explicitly. However, we still use the validation analysis of Section 4.1 and compare the resulting confidence interval of $\theta(x)$ with the true value of $\theta(x)$. We observe that Assumption 1 (and 4) holds for this problem instance.

We consider four candidate points: the optimal solution $\hat{x}$, a near-optimal point $x_1 =$ (10.0029, 9.4866, 9.0071, 8.5162, 7.9931, 7.5086, 7.0125, 6.4841, 5.9856, 5.5057, 4.9960, 4.5069, 4.0082, 3.5071, 3.0129, 2.5067, 2.0119, 1.4880, 0.9998, 0.4984)$'$ obtained by randomly perturbing $\hat{x}$, a further-away point $x_2 = (9.9, 9.4, 8.9, ..., 0.4)'$, and a relatively far-away point $x_3 = (9, 8.5, 8, ..., -0.5)'$.

Table 1 gives 95%-confidence intervals for $\theta(\hat{x})$ (column 2), $\theta(x_1)$ (column 3), $\theta(x_2)$ (column 4), and $\theta(x_3)$ (column 5) using (47) with varying sample size N. The last row gives the true values of $\theta(\hat{x})$, $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$. We observe that the confidence intervals cover the true value of the optimality function. When the value of the optimality function is some distance from zero, a tight confidence interval is obtained using a moderate sample size N. However, when the optimality function is close to zero, tightness can only be obtain by using a large sample size.

We also apply a hypothesis test based on a Chi-square statistic proposed in [35]. The test involves the null hypothesis that the current point satisfies the KKT conditions and the alternative hypothesis that they are not. For $\hat{x}$, we compute a p-value of 0.20 using a sample size of $N = 10^5$. Hence, with a test size of (for example) 0.05, we are unable to reject the null hypothesis. For x_1, x_2, and x_3, we compute p-values of essentially zero. Hence, in those cases we reject the null hypothesis with high confidence. While these conclusions are reasonable, they do not directly provide information about how "close" a candidate solution is to a Fritz-John point. In practice, we are rarely able to obtain a candidate solution that is a Fritz-John point. Hence, the "distance" to such a point becomes important. While [35] provides expressions for a confidence region for $\nabla f^0(x)$ that can be computed and compared with a user-defined region containing $0 \in \mathbb{R}^n$, it is more natural and convenient to condense $\nabla f^0(x)$ into a single number as achieved with the optimality function. As we see next (and in Section 4), the approach based on the optimality function also generalizes to constrained problems under mild assumptions.

6.2 Example 2: Validation Analysis for Deterministically Constrained Problem

The next problem instance generalizes a classical problem arising in search and detection applications. Consider an area of interest divided into n cells. A stationary target is located in one of the cells. A priori information gives that the probability that the target is in cell i is p_i, $i = 1, 2, ..., n$, with $\sum_{i=1}^n p_i = 1$. The goal is to optimally allocate T time units of search effort such that the probability of not detecting the target is minimized (see, e.g., p. 5-1 in [38]). We generalize this problem and consider a random search effectiveness in cell i per time unit and minimize the expected probability of not detecting the target. We let $x \in \mathbb{R}^n$, with x^i representing the number of time units allocated to cell i, and let $\omega = (\omega^1, \omega^2, ..., \omega^n)'$ be independent lognormally distributed random variables (with parameters $\xi^i = 100u^i$ and $\lambda^i = 0$, where $u^i \in (0, 1)$ are given data generated by independent sampling from a uniform distribution) representing the random search effectiveness in cell i. Then, the expected probability of not detecting the target is $f^0(x) = E[F^0(x, \omega)]$, where $F^0(x, \omega) = \sum_{i=1}^n p_i \exp(-\omega^i x^i)$. The decision variables are constrained by $\sum_{i=1}^n x^i \leq T$ and $x \geq 0$, where we use $T = 1$. We consider $n = 100$ cells. Assumption 1 (and 4) holds for this problem instance.

We consider three candidate solutions: $x_1 \in \mathbb{R}^{100}$, which is nearly optimal, $x_2 = (1/100, 1/100, ..., 1/100)' \in \mathbb{R}^{100}$, and $x_3 = (1/50, 1/50, ..., 1/50)' \in \mathbb{R}^{100}$, which is infeasible. We verify using long

				Confidence Intervals		
Method	N	M	K	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(59)	10^3	30	-	$(-0.000630, 0]$	$(-0.007837, 0]$	$(-1.048609, 0]$
	10^4	30	-	$(-0.000050, 0]$	$(-0.007783, 0]$	$(-1.048554, 0]$
	10^5	100	-	$(-0.000006, 0]$	$(-0.007483, 0]$	$(-1.009602, 0]$
(70)	10^3	-	5	$(-0.000464, 0]$	$(-0.007497, 0]$	$(-0.993391, 0]$
	10^4	-	5	$(-0.000049, 0]$	$(-0.007359, 0]$	$(-0.993278, 0]$
	10^5	-	5	$(-0.000006, 0]$	$(-0.007365, 0]$	$(-0.993201, 0]$
"Exact"				$\approx 8 \cdot 10^{-7}$	≈ -0.00736	≈ -0.99318

Table 2: Example 2: 95%-confidence intervals for $\theta(x_1)$ (column 3), $\theta(x_2)$ (column 4), and $\theta(x_3)$ (column 5) using (59) (rows 3-5) and (70) (rows 6-8) with varying sample size N and replications M and K. The last row gives approximate values of $\theta(x_1)$ (column 3), $\theta(x_2)$ (column 4), and $\theta(x_3)$ (column 5).

simulations ($N = 10^8$) that $\theta(x_1) \approx 8 \cdot 10^{-7}$, $\theta(x_2) \approx -0.00736$, and $\theta(x_3) \approx -0.99318$; see the last row of Table 2.

We consider both confidence interval (59) and (70). To compute (59), we first estimate μ by solving (14) using sample size N. Second, we compute $\overline{\eta}_{N,M}$ using that μ and M replications. In (70), we use $L = 1$ which leads to $K = 5$ when $\beta = 0.05$; see (68).

Table 2 provides 95%-confidence intervals for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (59) (rows 3-5) and (70) (rows 6-8) with varying sample size N and replications M and K. It appears that (59) tends to give slightly larger confidence intervals compared to (70) and that the computational effort is also greater. However, the use of (70) requires Assumption 3.

We confirm the confidence level stipulated by the confidence interval (70) by estimating coverage probabilities, i.e., the probability that the random confidence interval (70) includes $\theta(x)$. We find that 99%, 98% and 99% of 1000 (200 in the case of $N = 10^5$) independent replications of (70) cover $\theta(x_1)$ for $N = 10^3$, $N = 10^4$, and $N = 10^5$, respectively. Similar calculations for $\theta(x_2)$ and $\theta(x_3)$ result in coverage percentages of at least 97%. All these percentages are well above the stipulated 95%.

We also apply the hypothesis test of [35] and find a p-value of 0.65 for the case with x_1. Hence, we are unable to reject the null hypothesis that x_1 is a KKT point using any reasonable test size. In the case of x_2 and x_3, the p-values are essentially zero and the null hypothesis is rejected even with a small test size. As discussed in Section 6.1, we believe that results of the kind presented in Table 2 are more informative than such hypothesis tests.

6.3 Example 3: Validation Analysis for Problem with Expectation Constraint

We next consider an engineering design problem where the cost of a short structural column needs to be minimized subject to constraints on the failure probability and the aspect ratio; see [28]. The design variables are the width x^1 and depth x^2 of the column. In [29], we find that the failure

				Confidence Intervals		
Method	N	M	K	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(59)	10^3	30	-	$(-0.0554, 0]$	$(-0.7856, 0]$	$(-10.0301, 0]$
	10^4	30	-	$(-0.0074, 0]$	$(-0.8179, 0]$	$(-10.1692, 0]$
	10^5	100	-	$(-0.0014, 0]$	$(-0.7816, 0]$	$(-9.8631, 0]$
(70)	10^3	30	5	$(-0.0595, 0]$	$(-0.8129, 0]$	$(-10.6630, 0]$
	10^4	30	5	$(-0.0031, 0]$	$(-0.8229, 0]$	$(-10.1777, 0]$
	10^5	30	5	$(-0.0003, 0]$	$(-0.8137, 0]$	$(-10.3143, 0]$

Table 3: Example 3: 90%-confidence intervals for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (59) (rows 3-5) and (70) (rows 6-8) with varying sample size N and replications M and K.

probability for design $x = (x^1, x^2)$ can be approximated with high-precision by the expression $E[1 - \chi_4^2(r^2(x, \omega))]$, where ω is a four-dimensional standard normal random vector modeling random loads and material property, $\chi_4^2(\cdot)$ is the cumulative distribution function of a Chi-squared distributed random variable with four degrees of freedom, and $r(x, \omega)$ is the minimum distance from $0 \in \mathbb{R}^4$ to a limit-state surface describing the performance of the column given design x and realization ω; see [28, 29]. The failure probability is constrained to be no greater than 0.00135. Hence, we set $f^1(x) = E[1 - \chi_4^2(r^2(x, \omega))]/0.00135 - 1$. As in [28], we adopt the objective function $f^0(x) = x^1x^2$ and the additional constraints $f^2(x) = -x^1$, $f^3(x) = -x^1$, $f^4(x) = x^1/x^2 - 2$, and $f^5(x) = 0.5 - x^2/x^1$. In view of results in [29], Assumption 1 holds for this problem instance.

We consider three candidate designs: $x_1 = (0.334, 0.586)'$ is the best point reported in [28]; $x_2 = (0.346, 0.553)'$ is an infeasible solution reported in [28], and $x_3 = (0.586, 0.334)'$ is the "mirror image" of x_1. Table 3 presents similar confidence intervals as in Table 2, but with $\alpha = 0.1$ in (59) and $\alpha = \beta = 0.05$ in (70). The methods give comparable results. As observed earlier, a near optimal solution may require a relatively large sample size to ensure a tight confidence interval.

6.4 Example 4: Optimization and Validation Analysis for Full Problem

We next illustrate Algorithm 1 by considering an extension of Example 1. Let $F^0(\cdot, \cdot)$ be as defined in that example (see (73)) and also define $F^1(\cdot, \cdot)$ and $F^2(\cdot, \cdot)$ similarly, but with a^i and b^i being randomly and independently generated from a uniform distribution supported on $[0, 10]$ and $[0, 2]$, respectively. Moreover, we subtract 100 from these expression to construct constraints of the form $E[\sum_{i=1}^{20} a^i(x^i - b^i\omega^i)^2 - 100] \leq 0$. Hence, the resulting instance of P involves 20 decision variables, 60 independent random variables with uniform distribution each supported on $[0, 1]$, an expected value objective function, and two expected value constraint functions. Assumption 4 holds for this problem instance.

We apply Algorithm 1 to this problem instance using $x_0 = 0 \in \mathbb{R}^{20}$, $N_0 = 100$, $\Delta(N) = 1/\sqrt{N}$, and $\epsilon = \delta = 1$. Moreover, we let $k(N) = 2N$. The algorithm map $A_N(\cdot)$ is one iteration the Polak-He Phase 1-Phase 2 algorithm; see Section 2.6 in [25]. We refer to the iterations of Algorithm 1 with

Candidate Point	N	#iter.	Confidence Intervals $\psi(x_N^*)$	 $\theta(x_N^*)$	 $f^0(x_N^*)$
x_0^*	100	-	$(-\infty, -48.1472)$	$(-431.1261, 0]$	(5296, 5447)
x_{100}^*	100	302	$(-\infty, -2.0657)$	$(-8.9403, 0]$	(3411, 3533)
x_{200}^*	200	106	$(-\infty, -0.4903)$	$(-3.5880, 0]$	(3439, 3521)
x_{400}^*	400	104	$(-\infty, 0.5280)$	$(-2.0762, 0]$	(3419, 3477)
x_{800}^*	800	149	$(-\infty, 0.0672)$	$(-1.4028, 0]$	(3458, 3498)
x_{1600}^*	1600	66	$(-\infty, -0.0001)$	$(-0.7915, 0]$	(3453, 3482)
x_{3200}^*	3200	60	$(-\infty, -0.0107)$	$(-0.4043, 0]$	(3462, 3482)
x_{6400}^*	6400	75	$(-\infty, 0.0785)$	$(-0.2027, 0]$	(3466, 3481)
x_{12800}^*	12800	129	$(-\infty, 0.0125)$	$(-0.1082, 0]$	(3470, 3480)
x_{25600}^*	25600	79	$(-\infty, 0.0607)$	$(-0.1085, 0]$	(3467, 3474)
x_{51200}^*	51200	99	$(-\infty, 0.0499)$	$(-0.0609, 0]$	(3467, 3472)

Table 4: Example 4: 95%-confidence intervals for $\psi(x_N^*)$ (column 4) and $f^0(x_N^*)$ (column 6), and 90%-confidence intervals for $\theta(x_N^*)$ (column 5) for different candidate points generated by Algorithm 1.

the same sample size N as a stage. We run Algorithm 1 for ten stages and generate the candidate points x_0^*, x_{100}^*, x_{200}^*,..., x_{51200}^*. For each candidate point x_N^*, we compute the confidence intervals (51) and (70) using sample size $10N$ (1000 for x_0^*), replications $M = 30$ and $K = 23$, and $L = 1$. This selection of M, K, and L results in 95% confidence intervals for $\psi(x_N^*)$ and 90%-confidence intervals for $\theta(x_N^*)$.

Table 4 presents the confidence intervals for the candidate points. Columns 2 and 3 give the sample size and number of iterations used in each stage, respectively. Columns 4 and 5 give 95% confidence intervals for $\psi(x_N^*)$ and 90% confidence intervals for $\theta(x_N^*)$, respectively. We also compute 95% confidence intervals for $f^0(x_N^*)$ using the standard sample average estimators; see column 6. We see that x_N^* tends to become closer, as measured by $\theta(\cdot)$, to a Fritz-John point as the calculations progress. The ten stages required 6900 seconds of run time. The verification analysis needed 3300 seconds.

7 Conclusions

We have developed validation analysis procedures for nonlinear, possibly nonconvex, stochastic programs with expected value functions as both objective and constraint functions. The validation analysis assesses the quality of a candidate solution $x \in \mathbb{R}^n$ by its proximity to a Fritz-John stationary point as measured by the value of an optimality function at x. We construct an estimator of the optimality function and examine its consistency, bias, and asymptotic distribution. The estimator leads to confidence intervals for the value of the optimality function at x and, hence, confidence intervals for the "quality" of x. We also construct an implementable algorithm for solving smooth stochastic programs based on sample average approximations and a gradual increase in sample size.

Acknowledgement

The author acknowledges financial supported from Air Force Office of Scientific Research Young Investigator grant F1ATA08337G003.

Appendix

Proof of Proposition 3 Let $x \in X_\psi$. We only consider $x \neq \hat{x}$ since the result trivially holds when $x = \hat{x}$. We define $\tilde{\psi}(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ for any $x' \in \mathbb{R}^n$ by $\tilde{\psi}(x, x') \overset{\triangle}{=} \max\{f^0(x') - f^0(x), \psi(x')\}$. It follows by the mean value theorem and (9) that for any $x' \in \mathbb{R}^n$ and some $s^j \in [0,1]$, $j \in \mathbf{q}_0$,

$$\begin{aligned}
\tilde{\psi}(x, x') &= \max\Big\{\langle f^0(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^0(x + s^0(x' - x))(x' - x)\rangle, \\
&\qquad \max_{j \in \mathbf{q}}\{f^j(x) + \langle f^j(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^j(x + s^j(x' - x))(x' - x)\rangle\}\Big\} \\
&\leq \frac{1}{M}\max\Big\{\langle f^0(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2, \qquad (74) \\
&\qquad \max_{j \in \mathbf{q}}\{f^j(x) + \langle f^j(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2\}\Big\}.
\end{aligned}$$

Minimizing first the right-hand and then the left-hand side in (74) and using the fact that $\psi(x)_+ = 0$, we obtain that

$$\min_{x' \in \mathbb{R}^n} \tilde{\psi}(x, x') \leq \theta(x)/M. \qquad (75)$$

Using similar arguments, we also obtain that

$$\min_{x' \in \mathbb{R}^n} \tilde{\psi}(x, x') \geq \theta(x)/m. \qquad (76)$$

Let $\hat{x}' \in \mathbb{R}^n$ be the unique optimal solution of $\min_{x' \in \mathbb{R}^n} \tilde{\psi}(x, x')$. Since $\tilde{\psi}(x, x) = 0$, it follows that $\hat{x}' \in X_\psi$. From (75), we obtain that

$$\begin{aligned}
f^0(\hat{x}) - f^0(x) &= \min_{x' \in \mathbb{R}^n}\{f^0(x') - f^0(x) \mid \psi(x') \leq 0\} \\
&\leq \min_{x' \in \mathbb{R}^n}\{\tilde{\psi}(x, x') \mid \psi(x') \leq 0\} \\
&= \tilde{\psi}(x, \hat{x}') \\
&\leq \theta(x)/M,
\end{aligned}$$

which proves the right-most inequality in (10). We next prove the left-most inequality and consider three cases.

(i) Suppose that $\psi(\hat{x}') < \tilde{\psi}(x, \hat{x}')$ and $f^0(\hat{x}') - f^0(x) = \tilde{\psi}(x, \hat{x}')$. Then,

$$\min_{x' \in \mathbb{R}^n} \tilde{\psi}(x, \hat{x}') = \min_{x' \in \mathbb{R}^n}\{f^0(x') - f^0(x) \mid \psi(x') \leq 0\} = f^0(\hat{x}) - f^0(x). \qquad (77)$$

Hence, by(76), $\theta(x)/m \le f^0(\hat{x}) - f^0(x)$.

(ii) Suppose that $\psi(\hat{x}') = \tilde{\psi}(x,\hat{x}')$ and $f^0(\hat{x}') - f^0(x) = \tilde{\psi}(x,\hat{x}')$. We define $\hat{h} = \hat{x} - \hat{x}'$. Since $\hat{x}'$ is the unconstrained minimizer of $\tilde{\psi}(x,\cdot)$, it follows that the directional derivative of $\tilde{\psi}(x,\cdot)$ at $\hat{x}'$ is nonnegative in all directions, i.e.,

$$d\tilde{\psi}(x,\hat{x}';h) = \max\{\langle \nabla f^0(\hat{x}'), h\rangle,\ d\psi(\hat{x}',h)\} \ge 0, \tag{78}$$

for all $h \in \mathbb{R}^n$. By strict convexity of $f^0(\cdot)$,

$$\langle \nabla f^0(\hat{x}'), \hat{h}\rangle < (f^0(\hat{x}) - f^0(x)) - (f^0(\hat{x}') - f^0(x)) < 0. \tag{79}$$

Consequently,

$$d\psi(\hat{x}',\hat{h}) \ge 0. \tag{80}$$

Now, let $j' \in \hat{\mathbf{q}}(\hat{x}')$ be such that $d\psi(\hat{x}';\hat{h}) = \langle \nabla f^{j'}(\hat{x}'), \hat{h}\rangle$. Then, by the mean value theorem and (9) ,

$$f^{j'}(\hat{x}) \ge f^{j'}(\hat{x}') + \langle \nabla f^{j'}(\hat{x}'), \hat{h}\rangle + \tfrac{1}{2}m\|\hat{h}\|^2. \tag{81}$$

Hence, using (80) and (76), we obtain

$$\begin{aligned}\psi(\hat{x}) \ge f^{j'}(\hat{x}) \ &\ge\ \psi(\hat{x}') + d\psi(\hat{x}';\hat{h}) + \tfrac{1}{2}m\|\hat{h}\|^2 \\ &\ge\ \theta(x)/m + \tfrac{1}{2}m\|\hat{h}\|^2.\end{aligned} \tag{82}$$

Since $\psi(\hat{x}) \le 0$, we find that

$$\|\hat{h}\| \le \frac{\sqrt{2}}{m}\sqrt{-\theta(x)}. \tag{83}$$

In view of (9), X_ψ is compact. Hence, there exists a constant $c < \infty$ such that $\|\nabla f^0(x')\| \le c/4$ for all $x' \in X_\psi$. It now follows from (79) and (76) that

$$\begin{aligned}f^0(\hat{x}) - f^0(x) \ &>\ f^0(\hat{x}') - f^0(x) + \langle \nabla f^0(\hat{x}'), \hat{h}\rangle \\ &\ge\ \theta(x)/m - \|\nabla f^0(\hat{x}')\|\|\hat{h}\| \\ &\ge\ \frac{\theta(x) - c\sqrt{-\theta(x)}}{m}.\end{aligned} \tag{84}$$

(iii) Suppose that $\psi(\hat{x}') = \tilde{\psi}(x,\hat{x}')$ and $f^0(\hat{x}') - f^0(x) < \tilde{\psi}(x,\hat{x}')$. Then, due to the optimality of $\hat{x}'$ for $\tilde{\psi}(x,\cdot)$, $d\psi(\hat{x}', x' - \hat{x}') \ge 0$ for all $x' \in \mathbb{R}^n$. Using similar arguments as in (82), we obtain that for any $x' \in X_\psi$,

$$\begin{aligned}0 \ge \psi(x') \ &\ge\ \psi(\hat{x}') + d\psi(\hat{x}'; x' - \hat{x}') + \tfrac{1}{2}m\|x' - \hat{x}'\|^2 \\ &\ge\ \theta(x)/m + \tfrac{1}{2}m\|x' - \hat{x}'\|^2\end{aligned} \tag{85}$$

and

$$\|x' - \hat{x}'\| \le \frac{\sqrt{2}}{m}\sqrt{-\theta(x)}. \tag{86}$$

Hence,

$$\|\hat{x} - x\| \leq \|\hat{x} - \hat{x}'\| + \|x - \hat{x}'\| \leq 2\frac{\sqrt{2}}{m}\sqrt{-\theta(x)}. \tag{87}$$

It now follows from convexity of $f^0(\cdot)$ and (76) that

$$\begin{aligned} f^0(\hat{x}) - f^0(x) &> \langle \nabla f^0(x), \hat{x} - x \rangle \\ &\geq -\|\nabla f^0(x)\| \|\hat{x} - x\| \\ &\geq -\frac{c}{m}\sqrt{-\theta(x)}. \end{aligned} \tag{88}$$

The left-most inequality (10) now follows as a consequence of these three cases. □

References

[1] S. Alexander, T.F. Coleman, and Y. Li. Minimizing CVaR and VaR for a portfolio of derivatives. *J. Banking & Finance*, 30:583–605, 2006.

[2] F. Bastin, C. Cirillo, and P.L. Toint. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Mathematical Programming*, 108(2-3):207–234, 2006.

[3] G. Bayraksan and D.P. Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108:495–514, 2006.

[4] G. Bayraksan and D.P. Morton. Assessing solution quality in stochastic programs via sampling. In *Tutorials in Operations Research*, pages 102–122. INFORMS, 2009.

[5] G. Bayraksan and D.P. Morton. A sequential sampling procedure for stochastic programming. *Operations Research*, to appear, 2009.

[6] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer, New York, New York, 1990.

[7] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, 2. edition, 2003.

[8] P. Billingsley. *Probability and Measure.* Wiley, New York, New York, 1995.

[9] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer, 2000.

[10] Y. Ermoliev. Stochastic quasigradient methods. In *Numerical Techniques for Stochastic Optimization*, Yu. Ermoliev and R.J-B. Wets (Eds.), New York, New York, 1988. Springer.

[11] M. C. Fu. Gradient estimation. In *Simulation*, pages 575–616, Amsterdam, Netherlands, 2006. Elsevier.

[12] Gurkan, A. Ozge, and S. M. Robinson. Sample-path solution of stochastic variational inequalities. *Mathematical Programming*, 84(2):313–333, 1999.

[13] J. L. Higle and S. Sen. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming.* Springer, 1996.

[14] J.L. Higle and S. Sen. Statistical verification of optimality conditions for stochastic programs with recourse. *Annals of Operations Research*, 30:215–240, 1991.

[15] J.L. Higle and S. Sen. Duality and statistical tests of optimality for two stage stochastic programs. *Mathematical Programming*, 75:257–275, 1996.

[16] G. Infanger. *Planning under uncertainty: solving large-scale stochastic linear programs.* Thomson Learning, 1994.

[17] P. Kall and J. Meyer. *Stochastic Linear Programming, Models, Theory, and Computation.* Springer, 2005.

[18] S.H. Kim and B.L. Nelson. Selecting the best system. In *Simulation*, pages 501–534, Amsterdam, Netherlands, 2006. Elsevier.

[19] A. J. King and R. J. B. Wets. Epi-convergence of convex stochastic programs. *Stochastics and Stochastics Reports*, 34:83–92, 1991.

[20] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer, 2. edition, 2003.

[21] J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optimization*, 19:674–699, 2008.

[22] W. K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

[23] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization*, 19(4):1574–1609, 2009.

[24] V.I. Norkin, G.C. Pflug, and A. Ruszczynski. A branch and bound method for stochastic global optimization. *Mathematical Programming*, 83:425–450, 1998.

[25] E. Polak. *Optimization. Algorithms and consistent approximations.* Springer, New York, New York, 1997.

[26] S. M. Robinson. Sample-path optimization of convex stochastic performance functions. *Mathematics of Operations Research*, 21(3):513–528, 1996.

[27] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

[28] J. O. Royset, A. Der Kiureghian, and E. Polak. Reliability-based optimal design of series structural systems. *J. Engineering Mechanics*, 127(6):607–614, 2001.

[29] J. O. Royset and E. Polak. Extensions of stochastic optimization results from problems with simple to problems with complex failure probability functions. *J. Optimization. Theory and Application*, 133(1):1–18, 2007.

[30] L. Sakalauskas. Towards implementable nonlinear stochastic programming. In *Coping with Uncertainty*, pages 257–279. Springer, 2006.

[31] L. L. Sakalauskas. Nonlinear stochastic programming by Monte-Carlo estimators. *European J. Operational Research*, 137:558–573, 2002.

[32] A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

[33] A. Shapiro. Testing kkt conditions. Private Communication, June 2, 2003.

[34] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* Society of Industrial and Applied Mathematics, 2009.

[35] A. Shapiro and T. Homem-de-Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming*, 81:301–325, 1998.

[36] J. C. Spall. *Introduction to stochastic search and optimization.* John Wiley and Sons, New York, New York, 2003.

[37] S. W. Wallace and W. T. Ziemba. *Applications of Stochastic Programming.* Society for Industrial and Applied Mathematics, 2005.

[38] A. R. Washburn. *Search and Detection.* INFORMS, Linthicum, Maryland, 4. edition, 2002.

[39] H. Xu and D. Zhang. Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Mathematical Programming*, 119:371–401, 2009.